CR-185517

MICROCOMPUTER-BASED TESTS FOR REPEATED-MEASURES: METRIC PROPERTIES AND PREDICTIVE VALIDITIES

Robert S. Kennedy, Dennis R. Baltzley,
William P. Dunlap, Robert L. Wilkes,
and Lois A. Kuntz

Essex Corporation
1040 Woodcock Road, Suite 227
Orlando, FL 32803
(407) 894-5090

EOTR 89-02

1 May 1989

ACKNOWLEDGMENTS

Support for this work was under the following: National Aeronautics and Space Administration (Contract NAS9-17326), National Science Foundation (Grant 1S1-8521282), and U.S. Army Medical Research Acquisition Activity (Contract DAMD17-85-C-5095).

ABSTRACT

The present study is the fourth in a series to refine a menu of psychomotor and mental acuity tests. Field applications of such a battery are, for example, study of the effects of toxic agents or exotic environments on performance readiness, or determination of fitness for duty. The key requirement of these tasks is that they be suitable for repeated-measures applications, and so questions of stability and reliability are a continuing, central focus of this work. In the present study, after the initial (practice) session, seven replications of 14 microcomputer-based performance tests (32 measures) were completed by 37 subjects. Each test in the battery had previously been shown to stablize in less than five 90-second administrations and to possess retest reliabilities greater than $\underline{r}$ = 0.707 for three minutes of testing. However, all the tests had never been administered together as a battery and they had never been self-administered. In order to provide predictive validity for intelligence measurement, the Wechsler Adult Intelligence Scale-Revised (WAIS-R) and Wonderlic Personnel Test, measures of general intelligence, were obtained on the same subjects. In addition, a synthetic version of the Armed Services Vocational Aptitude Battery (ASVAB) was administered and American College Testing (ACT) scores were available for most subjects. The results showed that, in most cases, the 14 microcomputer tests achieved stability by Trial 3 or 4 for all the preferred measures. Instabilities, when they occurred (five tests, seven scores) were for the nonpreferred metric (percent correct, response latency). The tests all possess high test-retest reliability and low intersect correlations. Corrected-for-attenuation correlations imply a factorially diverse menu of tests.

Analyses indicated that the different global IQ measures correlated highly with each other (average $\underline{r}$ = .73). A "core" battery of eight microcomputer-based subtests was regressed on the traditional IQ measures and exhibited 21% to 65% common variance. Perhaps more importantly, they retained additional reliable variance which may be an index of factors in this battery which are not correlated with ordinary measures of intelligence. Finally, multiple correlations were examined between the IQ measures and performance measures at different stages of practice.

INTRODUCTION

This study is one of a series in which the collective goal is the development of a menu of tests embedded in a package of hardware and software to be used in repeated-measures studies of the effects of environmental and chemical stresses on human performance. In this work, tests are first subjected to an examination of their psychometric properties for repeated-measures testing (Bittner, Carter, Kennedy, Harbeson, & Krause, 1986; Kennedy, Dunlap, Jones, Lane, & Wilkes, 1985; Kennedy, Wilkes, Lane, & Homick, 1985). Repeated-measures testing is the most often employed design when studying changes due to environmental stress, drugs, and toxic substances, as well as disease-time-course effects. The primary psychometric qualities of tests which are to be employed in such repeated-measures designs are stability and reliability of between-subject variance. Also, the battery of tests must have a large number of alternate forms that are psychometrically equivalent. It is further helpful if these properties are achievable with an economy of time.

It is not uncommon for the development of test batteries to follow from cognitive theories (e.g., Hunter, 1975; Gullion & Eckerman, 1986; Wickens, Sandry, & Vidlich, 1983). When this is done, however, the theory is usually modified by new experience; these changes are often reflected in the test battery and, as a result, subsequent test evaluation and development efforts are seldom repeated. Thus, it becomes difficult or impossible to "mark" or "index" findings from early studies to different treatments or dosages which may be collected later.

To avoid this dilemma, our approach follows from classical test theory and it uses test theory as an engineering strategy to build a battery from parts. For example, test theory (Allen & Yen, 1979) makes simplifying assumptions such as that Obtained scores are comprised of a True score (T) and an Error score (E) regardless of the context of what they might measure. Test theory further assumes that True scores and Error scores are additive (rather than some other relationship), and that the True score portion of an Obtained score will be correlated with the True score portion when tested again, whereas the Error portion will not because it is nonsystematic or random. If fatigue occurs or learning is still going on (which can occur over repeated administrations of tests) then, in addition to the True score, there are other elements being measured which differ systematically from (i.e., are correlated with) ability on the test. In this case, the "True" score has two systematic parts and the assumptions of the theory are compromised. Such a theory therefore can accommodate hypothetical constructs like "controlled vs. automatic" processing (Ackerman & Schneider, 1984) or "components" (Sternberg, 1977) as they emerge. Only when a test is stable (i.e., systematic differences in automaticity, learning, or fatigue are no longer present) may the effects of treatments or agents be interpreted unambiguously. When individual differences are present which are not Error, then the retest correlation is proportional to the ratio of the True score to the total score variance. Therefore, a critical requirement of tests employed in repeated measures applications and within-subject designs, is that the tests be stable, and that alternate forms of the tests be parallel. The requirement for parallel forms is logically necessary for proper interpretation of any loss

(or gain) in the capacity or the performance being measured as being due to a treatment.

We believe that in the past when test batteries have been developed, little attention has been paid to certain areas of test theory, particularly stability. Relatedly, we have sought to determine if tests were also reliable. The criterion used in the current series of research has been a retest reliability of $r > .70$ for three minutes of testing (Kennedy, Wilkes, Dunlap, & Kuntz, 1987). Tests which are not reliable lack statistical power and so may be insensitive. Our tertiary purpose has been related to an explanation of the factorial diversity of tests. In previous research (e.g., Kennedy, Wilkes, Lane, & Homick, 1985), only four to six tests were studied. Recently, in this program, 20 additional tests have been examined over repeated measures (Kennedy, Wilkes, Kuntz, & Baltzley, 1988). In a similar effort, Englund, Reeves, Shingledecker, Thorne, Wilson, and Hegge (1987) described 25 additional prospective tests. Therefore, the current research plan was to administer as many tests as feasible at one time in order to provide information related to factorial diversity of those tests which were stable.

In addition, we sought to address the important issue of validity because the cardinal requirement of any test or test battery is that it be valid. The manual of standards and practices for tests (American Psychological Association, 1982) suggests that "good" tests need more than one kind of validity. Elsewhere, we have described content, construct, and to some extent face validity for tests in this battery (Kennedy, Carter, & Bittner, 1980), and we have reported sensitivity to stressors for some tests (Kennedy, Lane, & Kuntz, 1987). Although Hunt and Pellegrino, (1986), eschew predictive validity as a goal in itself, we believe such knowledge can guide the development of theory and the interpretation of tests. Because such a large literature exists relating scores on holistic measures of intelligence (or IQ) to most forms of academic and job performance, an attempt was made to link the microcomputer tests to measures such as American College Testing (ACT) Test, Armed Services Vocational Aptitude Battery (ASVAB), WAIS-R, and Wonderlic.

The last purpose in this study was to determine whether performance on the tests would be adversely affected if the computerized battery were self-administered and for the most part unproctored.

In summary, this study had five purposes: 1) to examine the metric properties (stability, reliability, and factor diversity) of 14 tests (32 scores). 2) to determine whether the tests could be self-administered; 3) to determine their predictive validity for global measures of intelligence; 4) to compare stability of their relations over practice and 5) to demonstrate that the battery would maintain its psychometric quality and validity even though self-administered and unproctored.

METHOD

SUBJECTS

The research subjects were obtained from undergraduate psychology classes at Casper College in Wyoming. Prior to subject solicitation, the Casper

College Human Use Committee reviewed and approved the purpose, methods, and procedures of the study. Sixty-four students indicated an interest in participating and a pool of potential subjects was established. The subject pool was based in part on availability of ACT scores and personal schedule conducive to group testing. Data collection was conducted in accordance with established guidelines for research with human participants (American Psychological Association, 1982). Initially 45 subjects were randomly selected from the pool for participation. During data collection four (4) subjects attrited the study for personal reasons and four (4) others were removed for lack of compliance with the established research procedures. The final sample consisted of 26 women and 11 men (i.e., total N = 37). The subjects ranged in age from 18 to 38 and were in good physical and mental health and represented freshmen to junior academic standings. The subjects were paid $4.50 per hour for their participation and motivation appeared to remain high over the 13 hours (approximate) of study obligations.

PROCEDURE

Prior to data collection, subjects received an introduction to the purpose of the study and were advised of the general research procedures. Subjects were directed to work quickly, accurately, and to the best of their abilities. Attempts to raise motivation and reduce test anxiety were made by indicating that the test batteries were the focus of the study as opposed to the subjects themselves. In our judgment, the subjects were motivated to perform, and were not adversely affected by performance anxiety.

Subjects were first tested with several standard paper-and-pencil measures of mental ability. These measures included the Wonderlic Personnel Test (Wonderlic, 1983), and a nonauthorized, synthetic Armed Services Vocational Aptitude Battery (ASVAB) (Steinberg, 1986). The ASVAB testing (3-hour administration time) and four replications of the Wonderlic (1-hour administration time) were conducted under group testing conditions. The ASVAB scores were obtained first, followed on a separate occasion by the Wonderlic testing. Scores for the standardized American College Testing (ACT) test (American College Testing Program, 1985) were obtained, with subjects' permission, through existing college files.

Prior to testing with the microcomputer-based battery, subjects were given a thorough introduction to the use of the self-administered testing system within a monitored classroom. They were encouraged to ask questions and to resolve difficulties. Testing procedures were reviewed, personal testing schedules were established, and handouts concerning procedure and scheduling were provided to each subject. Subjects were required to complete seven replications of the battery within a three-week period with multiple battery replications on a single test day not permitted. All self-testing was conducted within controlled laboratory rooms reserved for data collection associated with this study. Subjects were encouraged to self-test on an every-other-day basis (personal schedule permitting) and if more than seven days transpired between replications of the battery, an abbreviated "warm-up" practice battery was required. This occurred on 4% of the sessions (11/259). Twenty-five percent (10 subjects) were randomly assigned to a Zenith 181 lap-top computer and 75% (30 subjects) were randomly assigned to NEC PC 8210A portable computers. The superior memory capabilities of the Zenith PC,

permitted subjects assigned to the Zenith system to receive all their subtests on one computer. Subjects tested with the NEC system used two systems in tandem. Subtest order, practice, feedback, testing time, and instructions were held constant within both microcomputer-based testing systems. Random assignment of subjects to the two microprocessors facilitated the field testing of the Zenith 181 and provided for direct comparison of the two self-administered microcomputer-based testing systems. The focus of the current study concerns those subjects who used the NEC system. Previous field testing with this portable testing system, the NEC PC 8210A, has been carried out successfully (Kennedy, Wilkes, Lane, & Homick, 1985; Wilkes, Kennedy, Dunlap, & Lane, 1986).

Subject study obligations were concluded with the administration of a WAIS-R. The WAIS-R testing (approximately 1.25 hours of administration time) was conducted under laboratory conditions by a qualified psychometrist. The WAIS-R testing signaled the completion of all research obligations and qualified a subject for final payment.

MATERIALS

Criterion Mental Tests. Four different global aptitude/ability paper-and-pencil measures were employed in the study.

(1) The American College Testing (ACT) scores were obtained under the auspices of Casper College, Casper, Wyoming through existing college files. The ACT provides ability subscale scores in English, Math, Social Science, and Science as well as an overall composite score (American College Testing Program, 1985). The ACT is used by institutions of higher learning for prediction, advising, and placement purposes. While the Composite Score is regarded as a good indicator of general intelligence, including both verbal and quantitative components, the test also indexes high school and college achievement.

(2) The WAIS-R (The Psychological Corporation, 1981) provides both Verbal and Performance subscale scores and is one of the most widely used indicators of general intelligence. In clinical settings the test is also used as a diagnostic aid for disorders associated with brain damage and learning disabilities.

(3) Four forms of the Wonderlic Personnel Test (Wonderlic, 1983) were administered to each subject. The forms (I, II, IV & B) have been equated for comparability and each is administered in 12 minutes of testing. The Wonderlic is used in business and industry for personnel selection and placement and has normative data available for various occupations and educational levels. The Wonderlic is advertised as measuring "ability to learn" and is regarded as a short-form measure of general intelligence, however, it does not provide subscale measures of verbal and quantitative abilities.

(4) The synthetic Armed Services Vocational Aptitude Battery (ASVAB) was obtained from a book of facsimile tests (Steinberg, 1986) widely available in bookstores throughout the continental United States and used to practice for the ASVAB. It was compiled by a civil servant (Steinberg 1986) associated

with ASVAB testing for several years. As in the original ASVAB, this battery is composed of 10 subtests; General Science, Arithmetic Reasoning, Word Knowledge, Paragraph Comprehension, Numerical Operations, Coding Speed, Auto & Shop Information, Mathematics Knowledge, Mechanical Comprehension, and Electronic Information. Like the true ASVAB, the test may be administered in 144 minutes, however, instructions and procedures significantly increase the total testing time. One combination of ASVAB subtest scores serves as the Armed Forces Quantifying Test (AFQT) which determines acceptance into a particular branch of the armed services. Other scores are also derived from the ASVAB and serve to identify aptitude and training placement. The true ASVAB test is regarded as a measure of general intelligence with both verbal and quantitative components. There are no known normative psychometric data for the facsimile test employed.

Microcomputer based Assessment. The preliminary battery used was the Automated Performance Test System (APTS). Tests were selected which had previously exhibited stability and reliability. Collectively, the test evaluation and development efforts have been identified as the Automated Performance Test System (APTS), and are more fully described in Kennedy, Wilkes, Dunlap, and Kuntz (1987). The APTS program has been guided, in part, by earlier empirical findings of the Performance Evaluation Test for Environmental Research (PETER) program (cf. Bittner et al., 1986). The APTS is comprised of three subsystems: (1) hardware, (2) test programs, and (3) system control. Tests developed for the APTS were from a set of 30 performance measures found to be most statistically suitable for repeated-measures applications.

The Unified Tri-Service Committee Performance Assessment Battery (UTC-PAB) is similar to the Performance Assessment Battery (PAB) developed by the Walter Reed Army Institute of Research (WRAIR) (Thorne, Genser, Sing, & Hegge, 1985), but also contains tests from Navy (Naitoh, 1982) and Air Force sources (Shingledecker, 1984), and is composed of a variety of subtests which measure varying degrees of cognitive and visual-motor processing abilities. Test selection was by a tri-service committee of behavioral scientists. To our knowledge, the entire UTC-PAB tests have not yet been subjected to repeated-measures evaluation research. An abbreviated battery has been reported for a 10-day study with U.S. Navy pilots (Reeves & Thorne, 1988).

To varying degrees the formal selection of PAB and APTS subtasks for this study was based on the following considerations: (1) demonstrated conformity to general criteria for "good" performance tests (see Table 1); (2) potential for improved metric qualities given revised methods of application (see Bittner, Smith, Kennedy, Staley, & Harbeson, 1985); (3) indications representing well-differentiated factors associated with such cognitive processes and abilities as information processing, decision making, perception, and mental workload capacity; (4) present or potential compatibility with the microcomputer testing mode. Beyond these general considerations specific selection criteria were also applied to each candidate test. These criteria are discussed in Turnage, Kennedy, and Osteen (1987) and evaluate areas such as how much information is available, is it copywrited, how much does it cost, is instruction time reasonable, is feedback available, is special hardware necessary, and approximately 10 other related questions.

TABLE 1. DESCRIPTIONS OF TASK SELECTION CONSIDERATIONS

Selection Consideration	Descriptions
FACTOR	The factor(s) assessed by the measure as identified in the literature.
DOMAIN	The characteristics of the domain(s) of assessment of the capability of cognitive, perceptual, or motor skills.
TESTING MODE	The task mode or modes of administration identified as paper-and-pencil, microbased, or both.
TIME TO STABLE Xs AND SD	The total amount of elapsed time (massed or distributed) required for task mean and standard deviation stabilization for paper-and-pencil and/or microbased testing mode.
TIME TO DIFFERENTIAL STABILITY	The total amount of elapsed time (massed or distributed) required for task intertrial correlation stabilization for paper-and-pencil and/or microbased testing mode.
TASK DEFINITION	The reliability (r) of the task following the occurrence of differential stabilization for paper-and-pencil and/or microbased testing mode.
RELIABILITY EFFICIENCIES	The reliability (r) of a stabilized task standardized to a 3-minute administration base for paper-and-pencil and/or microbased testing mode.
EVALUATION CATEGORY	A global judgment of the acceptability of a paper-and-pencil and/or microbased test for use in repeated-measures research. Tasks are judged as recommended, acceptable-but-redundant, marginal, or unacceptable.
EVALUATION REFERENCE	The relevant study of stability and the original source of the measure.

Application of the criteria resulted in the selection of four PAB and 10 APTS tests (total battery included 14 subtests) for microcomputer-based adaptation and repeated-measures evaluation. All the tasks were timed and software programming ensured that comparable (i.e., parallel) but different forms were presented on repeated occasions of testing. Where appropriate, the tasks were scored for the number of items correctly answered (number correct), the percentage of items correctly answered (percent correct), and the average

time to respond (average response latency). The subtasks in order of appearance in the battery appear in Table 2. A brief description of each subtask is provided below:

TABLE 2. MICROCOMPUTER SUBTESTS, SOURCE, SUBTEST ORDER, TIME, PRACTICE AND FEEDBACL INFORMATION, AND TOTAL BATTERY ADMINISTRATION TIME

Order of Tests	Source [a]	Practice Time	Feedback	Trial Time	Trials/ Battery	Total Battery Task Time Less Practice	Total Battery Task Time For 7 Replications
1. AC	A	none[b]	yes	300[c]	1	300	2100
2. PTAP	A	10	yes	10	2	20	140
3. PC	A	30	yes	180	1	180	1260
4. GR	A	30	yes	180	1	180	1260
5. CR	P	30	yes	180	1	180	1260
6. MP	P	30	yes	180	1	180	1260
7. MN	A	30	yes	180	1	180	1260
8. TTAP	A	10	yes	10	2	20	140
9. RT1	A	none[b]	no	180	1	180	1260
10. AM	A	none[b]	no	90	1	90	630
11. NC	A	30	yes	90	1	90	630
12. CS	P	30	yes	180	1	180	1260
13. MR	P	30	yes	180	1	180	1260
14. NTAP	A	10	yes	10	2	20	140
Total		270		1950	17	1980	13860

[a] A = Test from APTS; P = Test from UTC-PAB
[b] Practice and trial number are the same
[c] All time data reported in seconds

AC - Auditory Counting
PC - Pattern Comparison
CR - Continuous Recall
MN - Manikin
AM - Associative Memory
NC - Number Comparison
MR - Matrix Rotation
PTAP - Preferred Hand Tapping
GR - Grammatical Reasoning
MP - Mathmatical Processing
TTAP - Two Finger Tapping
RT1 - Reaction Time
CS - Code Substitution
NTAP - Nonpreferred Tapping

(1) Auditory Counting (AC). The Counting test (Jerison, 1955) is accomplished by monitoring the repeated occurrence of a particular auditory stimulus. This test requires vigilance skills and short-term memory. The number of channels monitored permits one to grade workload. The participant is required to count the number of times a tone occurs. There are three different tones identified as low, medium, and high. In the high demand version of the test, which was the test administered in this experiment, the participant must count separately each low, each middle, and each high tone, and press the corresponding arrow key for every fourth low, every fourth middle and every fourth high tone. The rate of presentation for each individual stimulus is varied at either eight, six, or five presentations per minute. In a previous study (Kennedy, Wilkes, Kuntz, & Baltzley, 1988), all three demands of the auditory counting were studied, but the high-demand version was most reliable. Performance is scored according to the number of correct four counts, the number of omissions, and the number of errors for each demand version. The Counting tests are best presented with automated testing and are described as coding and memory-type tasks.

(2), (3), and (4). Tapping (TAP) Series. Tapping tests for assessment of motor skills/performance may be placed throughout the test battery to serve as a check against interfering factors during battery administration (e.g., boredom). The participant is required to alternately press the indicated keys as fast as he or she can with two fingers of either the preferred (PTAP), nonpreferred (NTAP), or from both hands (TTAP). Performance is based on the number of alternate key presses made in the allotted time. Kennedy, Wilkes, Lane, and Homick (1985), described tapping as a psychomotor skill assessing factors common to both Aiming and Spoke. Tapping has also been highly recommended for inclusion in a repeated-measures microcomputer battery (Kennedy, Dunlap, Jones, Lane, & Wilkes, 1985; Wilkes, Kennedy, Dunlap, & Lane, 1986).

(5) Pattern Comparison (PC). The Pattern Comparison task (Klein & Armitage, 1979) is accomplished by examining two patterns of asterisks that are displayed on the screen simultaneously. The participant is required to determine if the patterns are the same or different and respond with the corresponding "S" or "D" key. Patterns are randomly generated with similar and different pairs presented in random order. According to Bittner, Carter, Kennedy, Harbeson, and Krause (1986), Pattern Comparison "assesses an integrative spatial function neuropsychologically associated with the right hemisphere." A review of Pattern Comparison studies (Bittner, Carter, Kennedy, Harbeson, & Krause, 1986) indicated that the task is acceptable for use in repeated-measures research. Recent field testing with a microcomputer adaptation of the task (Kennedy, Dunlap, Jones, Lane, & Wilkes, 1985; Kennedy, Wilkes, Lane, & Homick, 1985; Wilkes, Kennedy, Dunlap, & Lane, 1986) resulted in strong recommendations for inclusion of Pattern Comparison in repeated-measures microcomputer test batteries.

(6) Grammatical Reasoning (GR). The Grammatical Reasoning Test (Baddeley, 1968) requires the participant to read and comprehend a simple statement about the order of two letters, A and B. Five grammatical transformations of statements about the relationship between the letters or symbols are made. The five transformations are: (1) active versus passive construction, (2) true versus false statements, (3) affirmative versus

negative phrasing, (4) use of the verb "precedes" versus the verb "follows," and (5) A versus B mentioned first. There are 32 possible items arranged in random order. The subject's task is to respond "true" or "false," depending on the verity of each statement. Performance is scored according to the number of transformations correctly identified. Grammatical Reasoning is described as measuring "higher mental processes" with reasoning, logic, and verbal ability, important factors in test performance (Carter, Kennedy, & Bittner, 1981). According to Bittner, Carter, Kennedy, Harbeson, and Krause (1986), Grammatical Reasoning "assesses an analytic cognitive neuropsychological function associated with the left hemisphere." Previous studies with Grammatical Reasoning, identified in Bittner, Carter, Kennedy, Harbeson, and Krause (1986), have indicated that the task is acceptable for use in repeated-measures research. Recent field testing with a microcomputer version of the task (Kennedy, Wilkes, Lane, & Homick, 1985; Kennedy, Dunlap, Jones, Lane, & Wilkes, 1985; Wilkes, Kennedy, Dunlap, & Lane, 1986) have resulted in strong recommendations for inclusion of Grammatical Reasoning in repeated-measures microcomputer test batteries.

(7) Reaction Time-1 Choice (RT1). The Visual Reaction Time Test (Donders, 1969) involves the presentation of a visual stimulus and measurement of a response latency to the stimulus. The subject's task is to respond as quickly as possible with a key press to a simple visual stimulus. On this test the subject is required to attend and respond to only one stimulus versus multiple stimulus. A short tone precedes at a random interval to signal that a "change" in the status of the stimulus is about to occur. The participant observes the stimulus for the change and then presses the response key as quickly as possible. Simple reaction time has been described as a perceptual task responsive to environmental effects (Krause & Bittner, 1982) and has been recommended for repeated-measures research (Bittner, Carter, Kennedy, Harbeson, & Krause, 1986; Kennedy, Dunlap, Jones, Lane, & Wilkes, 1985).

(8) Associative Memory (AM). This is a memory test (Underwood, Boruch, & Malmi, 1977) that requires the participant to view five sets of three letter trigrams that are paired with the numbers 1 to 5 and to memorize this list. After an interval, successive trigrams are displayed and the participant is required to press the key of the number corresponding to that letter set. In previous research (Krause & Kennedy, 1980) this associative memory task, using percent correct score, was recommended for inclusion in a performance test battery for environmental research.

(9) Number Comparison (NC). The Number Comparison task (Ekstrom, French, Harman, & Dermen, 1976) involves the presentation and comparison of two sets of numbers. The subject's task is to compare the numbers and decide if they are the same or different. Numbers may range from 3 to 7 digits in length with the second number always having the same number of digits the first. Only one digit in the second set may be different from the first set of numbers. Number Comparison has been described as a perceptual task involving perceptual speed, a factor important to performance. Previous research with Number Comparison has indicated that the task is acceptable for repeated-measures research (Bittner, Carter, Krause, Kennedy, & Harbeson, 1983; Carter & Sbisa, 1982).

(10) Manikin (MK). This performance test (Benson & Gedye, 1963) involves the presentation of a simulated human figure in either a full-front or full-back facing position. The figure has two easily differentiated hand-held patterns. One of the two patterns is matched to a pattern appearing below the figure. The subject's task is to determine which hand of the figure holds the matching pattern and respond by pressing the appropriate microprocessor key. Pattern type, hand associated with the matching pattern and front-to-back figure orientation are randomly determined for each trial. The Manikin Test is a perceptual measure of spatial transformation of mental images and involves spatial ability (Carter & Woldstad, 1985). Bittner et al. (1986) recommended the use of the Manikin Test latency scores, and Carter and Woldstad (1985) identified the Manikin Test for inclusion in microcomputer repeated-measures batteries.

(11) Continuous Recall (CR) - PAB. The Continuous Recall test (Hunter, 1975) indexes the subject's ability to serially encode and store information under changing memory states. The subject is presented with two single digit numbers, with one appearing above the other. The numbers are displayed for 5 seconds, followed by two other single digits similarly displayed during a 5-second interval. The subject's task is to determine if the bottom number of the first set is the same or different from the top number of the second set, and to respond with an appropriate key press. The task is continuous from set to set with the bottom digit of the previous display always being compared to the top digit of the following display. The Continuous Recall test is a measure of short-term memory requiring subjects to accurately maintain, update and access a store of information on a continuous basis (UTC-PAB, Englund et al., 1987). The Continuous Recall test has not been previously evaluated for repeated-measures applicability.

(12) Mathematical Processing (MP) - PAB. Mathematical Processing (Shingledecker, 1984) is a test that examines arithmetical operations as well as value comparison of numeric stimuli. The participant performs 1 to 3 addition or subtraction operation(s) in a single presentation. These operations correspond to low, medium, and high demand conditions. Then a response is made indicating whether the total is greater or less than a prespecified value of 5 using the arrow keys. The problems are randomly generated using only numbers 1 through 9. There are response deadlines for the problems corresponding to the demand characteristic of the test. The low demand version was used in this experiment.

(13) Matrix Rotation (MR) - PAB. This test (Phillips, 1974) assesses spatial orientation and short-term memory. A series of 5x5 cell matrices that contain five illuminated cells per matrix are presented (singly). The participant compares successive displays to determine if they are the same ("S") or different ("D"). Matrices are considered alike if the same matrix is rotated either 90 degrees to the left or 90 degrees to the right from the previously displayed matrix. Two successive matrices are never presented in exactly the same orientation. The stimulus remains on the screen until the subject makes a response.

(14) Code Substitution Test (CS) - PAB. Adapted from a paper-and-pencil version of the test contained in the Wechsler Adult Intelligence Scale from Wechsler (1958), this test is designed to measure associative learning ability

and perceptual speed. A string of nine letters and nine digits (numbers) are displayed across the screen in an arrangement so that the digit string is immediately below the letter string. Letters and digits are randomly paired for each test and their order is randomly assigned in the coding string. A test letter is presented at the bottom of the screen below the coding strings. The participant is to indicate which digit corresponds to that test letter in the display strings. The letter and digit associates change at 10-second intervals.

APPARATUS

NEC PC 8201A. Microcomputer testing was conducted with 27 subjects and was implemented on the NEC PC8201A microprocessor using scoring programs from the Essex Corporation APTS. The NEC PC8201A is configured around an 80C85 microprocessor with 64K internal ROM containing Basic, TELCOM, and a TEXT EDITOR. RAM capacity may be expanded to 96K onboard, divided into three separate 32K banks. Visual displays are presented on an 8 line LCD with 40 characters per line. Memory may be transferred to 32K modules with independent power supplies for storage or mailing. The entire package is lightweight (3.8 lbs), compact (110 W x 40 H x 130 D mm), and fully portable with rechargeable nickel cadmium batteries permitting up to four hours of continuous operation. Table 3 lists the technical features of the system which are more fully described in NEC Home Electronics (1983) and Essex (1985).

TABLE 3. NEC PC 8201A TECHNICAL SPECIFICATIONS

FEATURES	SPECIFICATIONS
SIZE	30 CM (11 IN) X 22 CM (8.25 IN) X 6 CM (2.5 IN). 1.7 KG (3.8 LBS)
CPU	80C85 (CMOS VERSION OF 8085) WITH 2.4 MHZ CLOCK
ROM	32K (STANDARD) - 128K (OPTIONAL)
RAM	24K (STANDARD) - 96K (OPTIONAL)
KEYBOARD	67 STANDARD (10 FUNCTIONS, 4 CURSOR DIRECTIONAL AND 58 ADDITIONAL)
DISPLAY	19 CM (7.5 IN) X 5.0 CM (2.0 IN) WITH REVERSE VIDEO OPTION. MAY BE CONFIGURED AS EITHER A 240 X 62 ELEMENT MATRIX OR 40 CHARACTERS X 8 LINE DISPLAY
INTERFACES	1 PARALLEL (CENTRONICS COMPATIBLE) AND 3 SERIAL (RS232C AND 6 & 8 PIN BERG) JACKS
POWER SUPPLY	4 AA NONRECHARGEABLE BATTERIES, OR RECHARGEABLE NICKEL-CADMIUM PACK, OR AC ADAPTER 50/60 Hz @ 120 VAC, OR EXTERNAL BATTERY SYSTEMS (e.g., 8 AMP HR)

ANALYSES

Repeated-measures experimental designs provide statistical power by reducing the proportion of error to true score through within-subject replications. Test theory (Allen & Yen, 1979) assumes that if practice, mood, apparatus, etc., or other systematic influences are absent from the error portion of the obtained score one may better estimate true score. Therefore, we examine our data for anomalies prior to formal analyses. This process is accomplished by examining data from four to five subjects at a time. Using the measures of number correct, percent correct, and average response latency, the data are plotted over trials for each subject for each test. The graphic presentation of the three measures provides for efficient inspection of all the scores. Data anomalies are then visually identified for appropriate action. The five criteria suggested by Turnage, Kennedy, and Osteen (1987) were followed: (1) if the subject's mean percent correct score is less than 60% on a two-choice (true/false) test, drop the subject for that particular test; (2) if the subject completes less than 75% of a series of test trials, drop the subject from that particular series; otherwise, retain the completed test trials; (3) if the subject does not respond after more than the beginning three trials, drop the subject for that session; (4) if the subject responds appropriately and systematically for all but one trial of a test, substitute a value for that trial (as the anomaly is probably a hardware or software malfunction); (5) if a subject has a mean response latency in any session that is more than 100% different from the group mean, drop the subject from the session.

Application of these criteria did not result in any subjects' data being dropped. However, missing values were substituted for trial one for three different subjects on Recall, Math Processing, and Manikin due to computer hardware problems. Also, a NEC software error occasioned the first trial of Number Comparison to be the length of a training trial instead of a full trial and as a result the number correct measure for Number Comparison was ignored for Trial 1.

METRIC ISSUES

General. For each test the reviewed and edited scores for number correct, percent correct, and average response latency were assessed for repeated-measures stability. These scores were chosen for analyses over others, based on recent findings by Turnage, Kennedy, and Osteen (1987) and Carter and Wolstad (1985). First, group means and standard deviations of these scores were examined for stability. Second, intertrial correlations of these scores were evaluated for evidence of correlational stability (Jones, Kennedy, & Bittner, 1981). Tests scores failing to demonstrate mean, standard deviation, or correlational stability were dropped from further analyses. Third, for tests demonstrating stability, task definition (average retest reliability after stabilization) and reliability efficiency (average stabilized intertrial correlations normalized to a three-minute base by the Spearman-Brown correction for changed test length, Guilford, 1954) were determined. Fourth, the intercorrelations of all tests, using the average score of the stable trials was established for all three scores (i.e. number correct, percent correct, and average response latency). Fifth, the analysis of the intercorrelations was repeated applying the correction for attenuation

formula (Spearman, 1904). Lastly, all the analyses were summarized to provide for the individual evaluation and direct comparison of the subtests.

Stability. Repeated-measures studies of environmental influences on performance require stable measures if changes in the treatment (i.e., the environment) are to be meaningfully related to changes in performance (Jones, 1970a). Of particular concern is the fact that a subject's scores may differ significantly over time due to lack of practice. The Jones two-process theory of skill acquisition (Jones, 1970a, b) maintains that the advancement of a skill involves an acquisition phase in which persons improve at different rates, and a terminal phase, in which persons reach or approximate their individual limits. The theory further implies that when the terminal phase is reached, scores will cease to deviate, despite additional practice. Unless tests have been practiced to this point of differential stability, the determination of changes in scores due to practice or some other variable would be impossible. Therefore, a stable test implies that the same thing is being consistently measured and an unstable test implies the converse, and is logically equivalent to the requirement for "parallel" test forms of classical test theory (Allen & Yen, 1979). For example, in a study of the effects of a toxic substance, if scores on a performance test remained the same before or after exposure, and if the test were not differentially stable, it would not be possible to determine whether a decline in performance was masked by practice effects or whether there was no treatment effect. Only after differential stability is clearly and consistently established between subjects can the investigator place confidence in the adequacy of his measures.

In this study means were considered stable if they were level, asymptotic or showed zero rate of change of slope over sessions. Similarly, standard deviations were considered stable if constant over sessions. Correlations were evaluated by a new graphical method. The average correlation of each session with all other sessions was computed, i.e., the average correlation of each row of the correlation matrix, excluding the diagonal element. This was compared to the "off-diagonal average" defined as the average of the three correlations among a given session and the two following sessions (i.e., for the first stability point the average of r_{12}, r_{23}, and r_{34} is used). Stability (i.e., Differential Stability or Intertrial Correlational Stability) was said to occur after that session where high ($r > .707$) and level cumulative average correlations were obtained. Additionally, the off-diagonal average correlation plots should be parallel to the average correlations of a trial with all other trials.

Task Definition. Task Definition is the average reliability of the stabilized task (Jones, 1980). Task Definition is obtained by averaging stable intertrial correlations. Higher average reliability improves power in repeated-measures studies when variances are constant. The lower the error within a measure the greater the statistical likelihood that mean differences will be detected, provided variances are also well behaved across repeated measures. Therefore, tasks with low task definition are insensitive to such differences and are to be avoided. Because different tasks stabilize at different levels, task definition becomes an important criterion in task selection. Task definitions for different tests, however, cannot be directly compared without first standardizing tests for test length (i.e., reliability efficiency).

Reliability Efficiency. Test reliability is known to be influenced by test length (Guilford, 1954). Tests with longer administration times and/or more items maintain a reliability advantage over tests with shorter administration times and/or fewer items. Test length must be equalized before meaningful comparisons can be made. A useful tool for making relative judgments is the reliability-efficiency, or standardized reliability, of the test (Kennedy, Carter, & Bittner, 1980). Reliability efficiencies are computed by correcting the reliabilities of different tests to a common test length by use of the Spearman-Brown prophecy formula (Guilford, 1954, p. 354). Reliability-efficiency not only facilitates judgments concerning different tests, but also provides a means for comparing the sensitivity of one test with the sensitivity of another test.

Stabilization Time. The evaluation of highly transitory changes in performance may be necessary when studying the effects of various treatments, drugs, or environmental stress. We believe that good performance measures should quickly stabilize following short periods of practice without sacrificing metric qualities, and good performance measures should always be economical in terms of testing time. We propose that a task under consideration for environmental research must be represented in terms of the number of trials and/or the total amount of time necessary to establish stability. Stabilization time must be determined for the group means, standard deviations, and intertrial correlations (differential stability).

SUMMARY OF METRIC REQUIREMENTS

We have described the formal requirements for stability and reliability of repeated measures tests in greater detail in several places (Bittner et al., 1986; Kennedy, Wilkes, Dunlap, & Kuntz, 1987). Summarizing stability analyses, criteria for evaluating a test as "good" in the present study were group means should be level, asymptotic, or show zero rate of change in slope over trials. Standard deviations should be constant, or covary as a proportion of the mean over trials. Correlations should be constant over trials (i.e., $r_{ij} = r_{ik} = r_{iz}$). Reliability analyses (i.e., task definitions) required that a test provide test retest reliability greater than $\underline{r} = .707$ for three minutes of testing.

RESULTS

Descriptive statistics for the global measures of intelligence are provided in Table 4. WAIS-R scores were at approximately the 75th percentile for persons of equivalent age and approximately average for a college group. The ACT scores were also about average for a college population. Wonderlic mean scores appeared consistent with this relation, that is, a slightly better than average mean score. There were no comparable data for the ASVAB.

TABLE 4. CRITERION MENTAL TESTS

Tests[a]	N	Mean	SD	Skew	Kurtosis
ASVAB	37	265.4	17.9	-1.17	3.43
ACT	32	103.9	24.9	-0.69	0.39
WONLK	37	270.5	52.2	-0.51	1.09
WVER	37	109.3	12.5	0.21	-0.78
WPER	37	118.4	14.7	0.02	-0.26
WAIS	37	109.6	12.8	0.26	-0.70

[a] Criterion Mental Test Codes
ASVAB - Armed Services Vocational Aptitude Battery (summed composite)
ACT - American College Testing Program (composite score)
WONLK - Wonderlic Personnel Test (summed composite of four administrations)
WVER - Wechsler Adult Intelligence Scale - Revised (verbal score)
WPER - Wechsler Adult Intelligence Scale - Revised (performance score)
WAIS - Wechsler Adult Intelligence Scale - Revised (composite score)

Correlations among the global measures of intelligence may be found in Table 5. Most of these tests are highly correlated with each other and all correlations are significant ($p < .01$). Since the tests of Table 4 may be expected to possess retest reliabilities (not shown) of $r > 0.80$ or 0.90 for each of the tests, after correction for attenuation, the global measures can be expected to share more than 50% common variance.

TABLE 5. CORRELATIONS[a] AMONG IQ MEASURES

	ASVAB	ACT	WONLK	WVER	WPER	WAIS
ASVAB						
ACT	0.787[b]					
WONLK	0.723	0.782				
WVER	0.486	0.538	0.535			
WPER	0.722	0.796	0.801	0.863		
WAIS	0.635	0.720	0.698	0.860	0.938	

[a] Correlations are based on N=37 subjects, with ACT N=32 the exception
[b] All correlations are significant at $p < 0.01$

Means and standard deviations for the 14 microcomputer tests which were examined in this study (see Table 6) came from the 27 subjects who used the NEC PC8201A. Examples are shown of stable (Figure 1) and unstable (Figure 2) correlations according to stability analyses performed (see also Kennedy, Wilkes, Kuntz, & Baltzley, 1988). In this graphic analysis, when correlations are high and level over sessions (e.g., the Tapping tests), they are considered differentially stable. When the correlations are low and/or not level over sessions (e.g., Continuous Recall) they are considered unstable and/or unreliable.

TABLE 6. MEANS AND STANDARD DEVIATIONS (N=27)

	Trials						
	1	2	3	4	5	6	7
Subtests							
AC (NC*)	10	12	14	14	15	14	16
	(6**)	(6)	(5)	(6)	(6)	(6)	(4)
PTAP (N)	32	36	36	38	38	39	40
	(13)	(12)	(11)	(10)	(9)	(7)	(7)
PC (NC)	109	119	125	128	129	130	132
	(21)	(19)	(19)	(21)	(22)	(22)	(20)
GR (NC)	39	37	44	44	47	48	48
	(10)	(15)	(11)	(14)	(17)	(16)	(14)
CR (NC)	50	65	75	81	82	87	93
	(35)	(37)	(39)	(44)	(44)	(48)	(50)
MP (NC)	98	112	124	130	131	136	142
	(23)	(23)	(22)	(21)	(22)	(21)	(18)
MK (NC)	72	83	95	101	103	107	109
	(28)	(32)	(32)	(29)	(33)	(34)	(32)
TTAP (N)	38	39	41	40	41	41	42
	(10)	(11)	(9)	(9)	(7)	(7)	(8)
RT1(RL)	453	366	311	311	323	330	329
	(242)	(151)	(62)	(69)	(84)	(97)	(88)
AM (NC)	12	14	13	13	15	15	15
	(4)	(5)	(5)	(4)	(5)	(4)	(5)
NC (NC)	19	57	60	65	64	63	66
	(19)	(19)	(18)	(11)	(11)	(12)	(14)
CS (NC)	61	63	66	66	66	67	67
	(9)	(6)	(6)	(5)	(5)	(6)	(6)
MR (NC)	65	76	80	81	81	84	86
	(23)	(22)	(24)	(24)	(24)	(25)	(25)
NTAP (N)	31	33	34	34	35	35	35
	(10)	(10)	(10)	(9)	(8)	(9)	(9)

* Codes: (N)=Number of Hits, (NC)=Number Correct, (RL)=Response Latency
** Standard Deviations in Parentheses
NA=Not analyzed due to software error

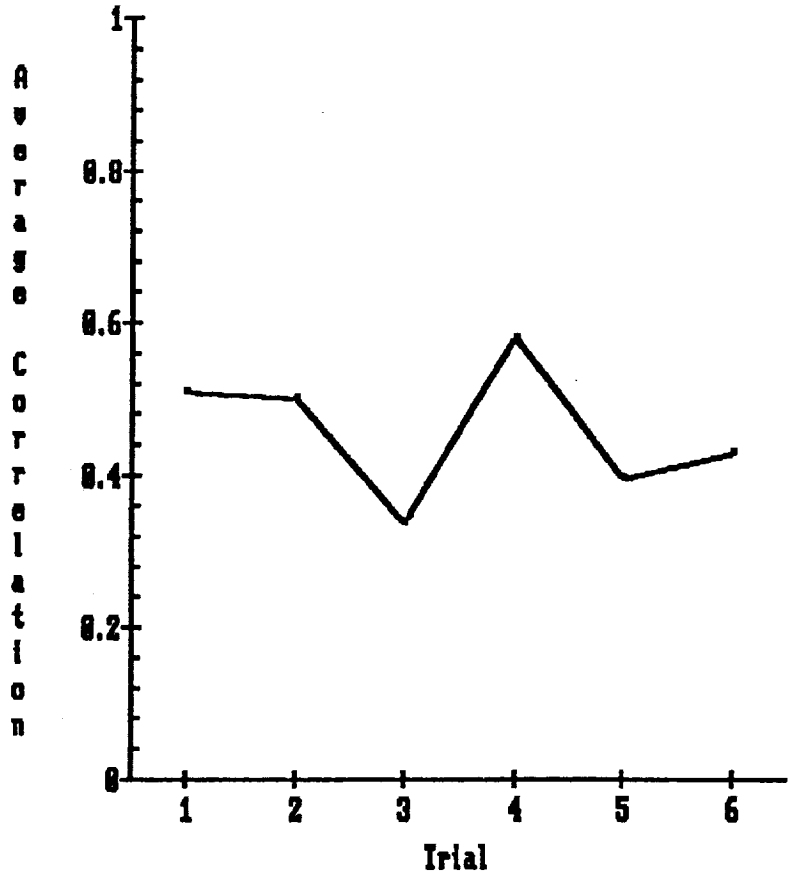

Figure 1. Average correlations (three adjacent sessions) over six trials for Continuous Recall.

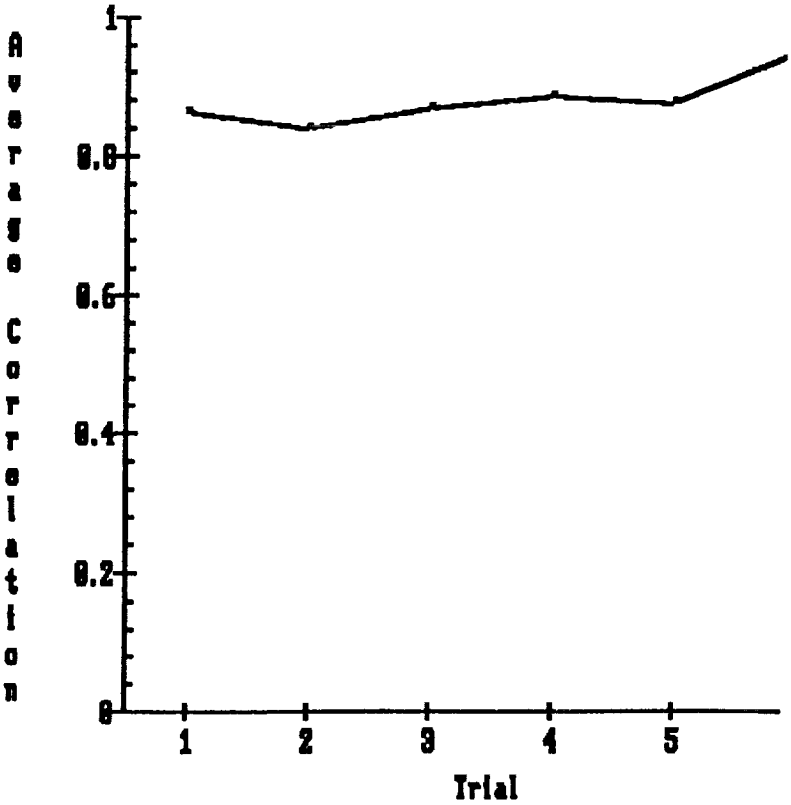

Figure 2. Average correlations (three adjacent sessions) over six trials for Tapping.

The data from Table 6 and Figures 1 and 2 are summarized in Table 7 where stability and reliability analyses appear for all tests across all scores. In general, test means revealed at least one score per subtest which stabilized between Trials 2 and 6, with standard deviations and intertrial correlations similarly well behaved. Only one test gave an indication of poor mean stability (i.e., Matrix Rotation for response latency). Three tests showed lack of homogenity of variance: Math Processing, Percent Correct and Response Latency; Associative Memory Response Latency; and Matrix Rotation, Response Latency. Task definitions were very high ranging from $\underline{r}$ = 0.78 for Code Substitution's Response Latency to $\underline{r}$ = 0.98 for Continuous Recall Number Correct and Preferred Hand Tapping. Using the Spearman Brown prediction for test length, stabilized reliability efficiencies were projected for a 3-minute test. These ranged from 0.78 for Code Substitution Response Latency to 0.99 for the Tapping task series. The Auditory Count task's value for reliability efficiency was lower than that obtained for task definition due to correcting the test down from a 5-minute to a 3-minute base. The consistently lower intertest reliability for percent correct scores is in agreement with previous research (Carter, Krause, & Harbeson, 1986; Dunlap, Kennedy, Harbeson, & Fowlkes, 1988) which has demonstrated that derived scores typically suffer from lower reliability when compared to directly measured scores. Latency scores may occasionally have higher task definitions but the advantage is usually in the second decimal place and latency scores also appear to possess higher between task correlations (Turnage et al., 1988) which may imply less factor density and diversity for the measures related to speed of response. Therefore, in the analyses which follow we have adopted Number Correct as the preferred metric except in those cases (viz., Reaction Time and Tapping) where other scores are warranted.

The findings from those tests in Table 7 which stabilized have been summarized into Table 8, where the preferred scores for each test have been shown along with the trial of stabilization.

Table 9 shows the stabilized retest reliability in the diagonal (in parentheses) for the 12 microcomputer tests (only one Tapping test included). A correlation matrix of the stabilized between-task correlations for the APTS appears above the diagonal in Table 9. Below the diagonal we have calculated corrected-for-attenuation values, as an index of overlap with other tests. It may be seen that the reliabilities of these tests are high (average $\underline{r}$ = .91) and even when corrected for attenuation (regardless of sign) the correlations among the 12 tests are only moderate $\underline{r}$ = .40, which implies a several-factor battery.

Cross-correlations between intelligence test score measures and the 14 microcomputer-based subtests are shown in Table 10. Virtually all of these are positive (Reaction Time, Response Latency, the exception) ranging from $\underline{r}$ = 0.04 to $\underline{r}$ = 0.81. The average $\underline{r}$'s for the microcomputer-based battery range from $\underline{r}$ = 0.10 to $\underline{r}$ = 0.66. Generally, the highest relationships are seen with the Wonderlic and the lowest with tests from the WAIS-R performance subtests.

TABLE 7. MICROBASED PERFORMANCE TESTS TRIAL OF STABILIZATION AND STABILITY INDICES BASED ON N = 27 FOR NUMBER CORRECT[a], PERCENT CORRECT[b], AND RESPONSE LATENCY SCORES[c]

		Trial of Stabilization					
Tests	Score Type	$\bar{X}$	SD	r	Total Task	Task Definition	Reliability Efficiency
AC	NC	3	1	2	3	.86	.79[d]
PTAP	N[e]	2	2	2	2	.98	.99
PC	NC	3	2	1	3	.92	.92
	PC[b]	2	3	4	4	.86	.86
	RL[c]	2	3	3	3	.81	.81
GR	NC	3	4	3	4	.94	.94
	PC	1	3	3	3	.89	.89
	RL	3	4	3	4	.93	.93
CR	NC	4	4	2	4	.98	.98
	PC	1	2	1	2	.88	.88
	RL	3	3	3	3	.96	.96
MP	NC	3	1	3	3	.93	.93
	PC	1	U	3	U	--	--
	RL	3	U	3	U	--	--
MK	NC	4	2	3	4	.97	.97
	PC	3	3	3	3	.95	.95
	RL	4	4	2	4	.94	.94
TTAP	N	1	3	1	3	.97	.99
RT1	RL	3	3	3	3	.86	.86
AM	NC	2	1	3	3	.88	.94
	PC	2	1	3	3	.88	.94
	RL	2	U	U	U	--	--
NC	NC	2	4	4	4	.91	.95
	PC	1	3	U	U	--	--
	RL	2	2	2	2	.87	.93
CS	NC	2	2	2	2	.85	.85
	PC	1	1	U	U	--	--
	RL	4	4	2	4	.78	.78
MR	NC	2	1	2	2	.90	.90
	PC	1	2	U	U	--	--
	RL	U	U	4	U	--	--
NTAP	N	1	1	1	1	.97	.99

a NC = Number Correct Score
b PC = Percent Correct Score
c RL = Response Latency Score
d Lower reliability efficiencies are reflected (in part) due to correcting test from 5-min. to 3-min. base.
e N = Total number of alternate key presses
U = Unstable

TABLE 8. SUMMARY RESULTS FOR PREFERRED SCORES FOR EACH TEST

Tests	Score Type	Trial of Stabilization $\bar{X}$	SD	r	Total Task	Task Definition	Reliability Efficiency
AC	NC[a]	3	1	2	3	.86	.79[b]
PTAP	N[c]	2	2	2	2	.98	.99
PC	NC	3	2	1	3	.92	.92
GR	NC	3	4	3	4	.94	.94
CR	NC	4	4	2	4	.98	.98
MP	NC	3	1	3	3	.93	.93
MK	NC	4	2	3	4	.97	.97
TTAP	N	1	3	1	3	.97	.99
RT1	RL[d]	3	3	2	3	.86	.86
AM	NC	2	1	3	3	.88	.94
NC	NC	2	4	4	4	.91	.95
CS	NC	2	2	2	2	.85	.85
MR	NC	2	1	2	2	.90	.90
NTAP	N	1	1	1	1	.97	.99

a NC = Number Correct Score
b Lower reliability efficiencies are reflected (in part) due to correcting test from 5-min. to 3-min. base.
c N = Total number of alternate key presses
d RL = Response Latency Score

TABLE 9. CROSS-TASK CORRELATIONS (ABOVE DIAGONAL) RELIABILITIES (IN PARENTHESES) CORRECTED FOR ATTENUATION VALUES (BELOW DIAGONAL) AMONG STABILIZED TRIALS

	AC	PCNC	GRNC	CRNC	MPNC	MKNC	RT1	AMNC	NCNC	CSNC	MRNC	NTAP
AC	(.86)	.36	.31	.17	.47*	.53*	-.47*	.20	.43	.46*	.09	.62**
PCNC	.40	(.92)	.46*	.39	.80**	.55*	-.69**	-.11	.61**	.62**	.41	.32
GRNC	.38	.49	(.94)	.34	.53*	.36	-.48*	-.20	.52*	.39	.12	.25
CRNC	.19	.41	.35	(.98)	.35	.37	-.19	.15	.26	.16	.59**	.16
MPNC	.53	.86	.57	.37	(.93)	.68**	-.62**	-.10	.77**	.64**	.31	.42
MKNC	.58	.58	.38	.38	.72	(.97)	-.27	-.05	.51*	.50*	.32	.43
RT1	-.55	-.77	-.53	-.21	-.70	-.30	(.86)	.14	-.58**	-.62**	-.05	-.34
AMNC	.23	-.12	-.22	.16	-.11	-.05	.16	(.88)	.05	.02	.09	.09
NCNC	.49	.67	.56	.28	.84	.54	-.65	.06	(.91)	.82**	.16	.43
CSNC	.53	.70	.44	.18	.72	.55	-.72	.02	.93	(.85)	.25	.53*
MRNC	.10	.45	.13	.63	.34	.34	-.06	.10	.18	.29	(.90)	.10
NTAP	.68	.34	.26	.16	.44	.44	-.37	.10	.46	.58	.11	(.97)

* $p < 0.05$
** $p < 0.01$

TABLE 10. CROSS-CORRELATIONS[a] BETWEEN IQ MEASURES AND MICROBASED SUBTESTS

	ACT	WONLK	ASVAB	WVER	WPER	WAIS	AVG r
AC	0.52**	0.60**	0.44*	0.35*	0.29	0.36*	0.43
PHT	0.55**	0.57**	0.53**	0.41*	0.42*	0.44*	0.49
PC	0.65**	0.68**	0.81**	0.63**	0.56**	0.64**	0.66
GR	0.52**	0.52**	0.53**	0.28	0.25	0.28	0.40
CR	0.41*	0.42*	0.41*	0.53**	0.17	0.43*	0.40
MP	0.62**	0.73**	0.81**	0.52**	0.36*	0.49**	0.59
MK	0.50**	0.66**	0.62**	0.42*	0.39*	0.44*	0.51
TTAP	0.20	0.21	0.19	0.04	0.25	0.12	0.17
RT1[b]	-0.50**	-0.65**	-0.66**	-0.40*	-0.40*	-0.42*	-0.57
AM	0.14	0.03	-0.10	0.22	0.10	0.21	0.10
NC	0.51**	0.65**	0.75**	0.33	0.47**	0.40*	0.52
CS	0.42*	0.59**	0.65**	0.30	0.50**	0.40*	0.47
MR	0.02	0.24	-0.25	0.43*	0.23	0.39*	0.18
NTAP	0.39*	0.35*	0.34*	0.18	0.15	0.17	0.26
Average r	0.35	0.40	0.36	0.30	0.27	0.31	

[a] Correlations are based on N=27, with ACT N=23 the exception
* $p < 0.05$
** $p < 0.01$
[b] Negative correlations for RT1 are due to scoring method (response latency)

Table 11 shows similar relationships for the individual subtests of the synthetic ASVAB used here. In general, the "information tasks" (e.g., Auto and Shop Information, Electronic Information) show lower overall correlations than the "ability" measures (Coding Speed, General Science, Mechanical Comprehension), the exception being Word Knowledge which may be more of an ability index than an information test. The Tapping series, Associative Memory, and the two spatial tests (Manikin and Matrix Rotation) show the lowest correlations with the ASVAB tests.

TABLE 11. CROSS-CORRELATIONS[a] BETWEEN ASVAB SUBTESTS AND MICROBASED SUBTESTS

	GS	AR	WK	PC	NO	CS	AS	MK	MC	EI
AC	0.21	0.22	0.38	0.16	0.16	0.36*	-0.06	0.28*	0.47**	0.20
PTAP	0.54**	0.23	0.58**	0.49**	0.12	0.39*	-0.33*	0.20	0.47**	0.39*
PC	0.63**	0.56**	0.56**	0.55**	0.42*	0.64**	-0.19	0.58**	0.52**	0.20
GR	0.32*	0.26	0.40*	0.47**	0.34*	0.40*	-0.02	0.44*	0.39*	0.15
CR	0.31	0.48**	0.32*	0.52**	0.21	0.13	-0.03	0.49**	0.22	0.33*
MP	0.52**	0.54**	0.53**	0.53**	0.40*	0.76**	-0.19	0.54**	0.36*	0.16
MK	0.34*	0.41*	0.53**	0.40*	0.05	0.59**	-0.08	0.29	0.25	0.24
TTAP	0.14	-0.15	0.41*	0.18	0.13	0.36*	-0.44*	-0.14	0.12	0.00
RTl	-0.46**	-0.45**	-0.53**	-0.38*	-0.68**	-0.51**	0.02	-0.44*	-0.42*	-0.13
AM	-0.03	0.03	-0.14	-0.07	-0.06	-0.30	0.28	-0.07	0.18	0.43*
NC	0.41*	0.43*	0.32*	0.35*	0.33*	0.78**	-0.07	0.40*	0.32*	0.18
CS	0.28	0.40*	0.37*	0.25	0.39*	0.71**	-0.12	0.37*	0.37*	-0.05
MR	-0.01	0.38*	0.07	0.14	0.16	0.20	-0.23	0.31	0.24	-0.05
NTAP	0.23	-0.004	0.46**	0.29	0.03	0.41*	-0.39*	0.09	0.18	0.17

[a] Correlations are based on N=27, with ACT N=23 the exception
* $p < 0.05$
** $p < 0.01$

GS=General Science
AR=Arithmetic Reasoning
WK=Word Knowledge
PC=Paragraph Comprehension
NO=Numerical Operations
CS=Coding Speed
AS=Auto and Shop Information
MK=Mathematics Knowledge
MC=Mechanical Comprehension
EI=Electronic Information

In order to examine the relationships between the microcomputer based tests and the various IQ (reference) tests via multiple regression, we first established a core battery of APTS tests based upon our prior experience with these tests in terms of their psychometric properties, their earlier demonstrated predictive power, and their factorial richness. This was necessary because multiple regression coefficients are remarkably vulnerable to shrinkage; thus, although the multiple R will continue to increase with the addition of more variables, the R corrected for shrinkage, the "adjusted" R, will decrease. The core battery selected was composed of the following eight tests: Pattern Comparison; Nonpreferred Hand Tapping; Code Substitution; Associative Memory; Simple Reaction Time; Grammatical Reasoning; Manikin; and

Matrix Rotation. As can be seen in Table 12, the relationships between the core battery and the various IQ or "g" measures are uniformly high. Furthermore, even after correction for shrinkage the correlations are still substantial, except in the case of the WAIS based measures, where although positive, the relations are at best moderate. It is important to note that the strongest relationships are with the ACT and ASVAB which are both general intelligence tests whose basic purposes are for selection.

TABLE 12. SQUARED MULTIPLE CORRELATIONS OF EACH IQ MEASURE PREDICTED BY THE MICROBASED BATTERY SUBTESTS

TESTS	R	R-SQUARED	ADJUSTED R	F	DF	P
ASVAB	.87	.75	.80	6.82	8/18	.000
ACT	.85	.72	.75	4.53	8/14	.007
WONLK	.84	.71	.76	5.59	8/18	.001
WVER	.74	.54	.57	2.52	8/17	.052
WPER	.63	.40	.34	1.40	8/17	.266
WAIS	.72	.52	.55	2.34	8/17	.067

In summary, the microcomputer-based tests correlate with holistic measures of intelligence, and possess sufficient reliability still in reserve in order to be potentially predictive of factors not presently measured by the intelligence-type tests.

A final analysis of these data involved assessing the relationship of the "core" battery of APTS subtests to general IQ or g measure at various stages of practice on the cognitive-performance tests. For purposes of this analysis, Replications 2 and 3 were considered early trials, Replications 4 and 5 to be middle trials, and Replications 6 and 7 to be trials late in practice. Multiple correlations of the "core" battery and the general IQ measures are shown in Table 13. As can be seen, the strength of the relationship between both the WAIS and the Wonderlic and the core battery decreased as practice proceeded. On the other hand, the correlations with the ASVAB and ACT scores did not appear to change dramatically as a function of practice.

TABLE 13. MULTIPLE CORRELATIONS BETWEEN THE CORE PERFORMANCE BATTERY AND REFERENCE TESTS OF INTELLECTUAL ABILITY AS FUNCTIONS OF PRACTICE

	Early	Mid	Late	Early	Mid	Late
	R	R	R	R_C	R_C	R_C
WAIS-R	.78	.75	.70	.65	.59	.50
Wonderlic	.89	.86	.81	.84	.79	.71
ASVAB	.86	.88	.83	.79	.84	.75
ACT	.85	.80	.84	.56	.44	.55

DISCUSSION

This study evaluated the metric properties of 14 mental acuity tests implemented on a portable microcomputer and compared them to established holistic measures of intelligence. Several of the tests were from the APTS battery and had previously been shown to be stable and reliable. Four of the 14 tests also appear in the UTC-PAB battery (Englund et al., 1987). In this study, preferred scores (usually Number Correct) for these tests generally stabilized quickly and with adequate reliabilities. Eighteen of the remaining scores included percent correct and latencies, seven of these were unstable, and they were generally less reliable than the Number Correct scores, repeating a finding reported previously (Turnage, Kennedy, & Osteen, 1988).

Correlations among the microcomputer-based tests were generally low, and given the high retest reliabilities of all the tests, it should be possible to create a multifactor battery of tests using the correlation matrix shown in Table 9 as a guide.

The most important result of this study is the addition of further evidence attesting to the excellent psychometric qualities exhibited by the 14 tests in a repeated-measures framework. All of the tests successfully passed stringent multiple hurdles for stability and reliability for their respective preferred scores. It is from this base that confident statements can be made with respect to subsequent issues such as determining factorial richness or interpreting the more complex interrelationships with global measures of IQ.

An additional point which should not be underestimated is the fact that all testing was self-administered. Other than preliminary orientation and practice sessions, the subjects were not directly supervised, yet excellent results were obtained. Due to the computer configuration (e.g., internal clocks and built-in security of the programs) any effort to test out-of-schedule or tamper with the apparatus was immediately obvious. Hence this research provides evidence for a new more flexible avenue in repeated-measures testing. The applications are many, such as robust testing in remote and/or hazardous areas where proctoring the testing process is not feasible.

Four different global measures of intelligence were intercorrelated and revealed considerable overlap. When the holistic measures of intelligence were compared to the microcomputer-based subtests the average r^2 varied from essentially zero ($\underline{r}^2$ = .0004 for Matrix Rotation and ACT) to $\underline{r}^2$ = .66 for Pattern Comparison and Math Processing with ASVAB. This finding is consistent with that of Hunt and Pelligrino (1986) and Detterman (1984) and implies that microcomputer-based tests are tapping factors available from more traditional paper-and-pencil and individually administered tests. However, the retest reliabilities of the microcomputer tests are so large (i.e., $\underline{r} > .707$) that it is evident there is considerable additional predictive power in the microcomputer tests.

The present study is one of a series where the collective, programmatic goal is development of a menu of tests implemented on a portable microcomputer(s) with excellent metric properties. Classical test theory (Allen & Yen, 1979) not cognitive theory (e.g., Carroll, 1974; Hunter, 1975)

is the guiding force. Of the 14 separate tests, all are stable except for Continuous Recall. All of the tests had acceptable levels of retest reliability, and except for one (simple Reaction Time) met or exceeded minimum requirements with the lowest reliability at 0.85 for Code Substitution.

The results of this study reveal that stable measures of performance implemented on a microcomputer test battery, bear a strong relationship to global measures of intelligence, such as the synthetic ASVAB, ACT, Wonderlic, and to a lesser extent, performance subtests of the Weschler Adult Intelligence Scale. Multiple correlational analyses of these microcomputer tests regressed against criterion scores on global measures of intelligence revealed 60%-87% total common variance (after adjustment 65%) for the synthetic ASVAB, 55% for the ACT, and 62% for the Wonderlic, and nearly 30% of the Weschler Adult Intelligence Scale verbal subtest. This experiment shows that in a short period of time (the microcomputer battery of selected tests only takes 15 minutes for each administration) it is possible to account for a substantial portion of variance in global measures of intelligence. Thus, over half of the variance of the much longer (2.4+ hours) ASVAB and ACT (3.5 hr.) can perhaps be predicted by this shorter microcomputer battery. Because predictive validity could be expected to increase as retest reliabilities increase, following the Spearman prophecy formula we theorize that a battery 2 to 3 times as long could add 10%-20% additional variance particularly if specific subtests were selected for emphasis. The best tests for all holistic measures are Pattern Comparison and Math Processing, with Reaction Time and Manikin a close third and fourth. Note that these tests do not have any obvious verbal content and so may be likely to be collectively "fair." Other tests may be more or less useful depending on which global measure is selected as may be extrapolated from the interrelations shown in Tables 10 and 11. For example, a battery of all the tests leaving out Matrix Rotation, Associative Memory, and Two Finger Tapping seems to be a good choice for Wonderlic.

The microcomputer battery is made up of tests which have correlations of around $\underline{r} = 0.2$ to $\underline{r} = 0.3$ between tests and as a result in combination can be expected to measure markedly different factors and constructs. Notably, their retest reliabilities tend to be greater than $\underline{r} = 0.707$ and in some cases exceed $r = 0.95$ for very brief (3 minutes each) periods of testing. Since the global measures also possess high retest reliabilities, this implies that the tests in the menu are measuring something which is shared with these global measures of intelligence, but also, that they are measuring something else. They therefore hold promise for being added to existing personnel measures such as the ASVAB for use as primary or secondary selection techniques.

The findings summarized in Table 13 regarding changes in correlation between the core battery and various IQ or "g" measure as functions of practice on the cognitive-performance battery are both interesting and important. The general tests for which the multiple correlation coefficients clearly dropped were the WAIS and Wonderlic which are tests thought to be relatively pure measures of IQ. The global tests that were more stable relative to stage of practice were the ASVAB and ACT, both of which are tests that are more slanted toward performance and achievement. Fleishman and Hempel (1954, 1955), among others, have studied the change in factorial structure of performance test as a function of stage of practice, and refer to the process underlying later trial skilled performance as the emergence of

"automaticity." Fleishman and Rich (1963) found parallel findings of correlations that drop with practice between reference tests of intellectual ability with what they termed skill development tasks. A recent summary of the history, data, and theory of the relationship of skilled performance to global measures of intellectual ability can be found in Ackerman (1987).

A major finding of the current study is that the correlations between the core battery and the ASVAB and ACT, the test used primarily for selective purposes, is both substantial and relatively stable. Since ASVAB has been shown to be related to military grades and job performance (Zeider, 1987) there are implied relevances of the APTS tests for military selection.

REFERENCES

Ackerman, P. L. (1987). Individual differences in skill learning: An integration of psychometric and information processing perspectives. Psychological Bulletin, 102, 3-27.

Ackerman, P. L., & Schneider, W. (1984, August). Individual differences in automatic and controlled information processing (Rep. No. HARL-ONR-8401). Champaign, IL: Human Attention Research Laboratory.

Allen, M. J., & Yen, W. M. (1979). Introduction to measurement theory. Monterey, CA: Brooks Cole.

American College Testing Program (1985). The ACT. Iowa City, IA: The American College Testing Program.

American Psychological Association (1982). Ethical principles in the conduct of research with human participants. Washington, DC: American Psychological Association.

Baddeley, A. D. (1968). A three-minute reasoning test based on grammatical transformation. Psychonomic Science, 10, 342-***.

Benson, A. J., & Gedye, J. L. (1963). Logical processes in the resolution of orientation conflict (Report 259). Farnborough, UK: Royal Air Force, Institute of Aviation Medicine.

Bittner, A. C., Jr., Carter, R. C., Kennedy, R. S., Harbeson, M. M., & Krause, M. (1986). Performance Evaluation Tests for Environmental Research (PETER): Evaluation of 114 measures. Perceptual and Motor Skills, 63, 683-708.

Bittner, A. C., Jr., Carter, R. C., Krause, M., Kennedy, R. S., & Harbeson, M. M. (1983). Performance Evaluation Tests for Environmental Research (PETER): Moran and computer batteries. Aviation, Space, and Environmental Medicine, 54, 923-928.

Bittner, A. C., Jr., Smith, M. G., Kennedy, R. S., Staley, C. F., & Harbeson, M. M. (1985). Automated portable test system (APTS): Overview and prospects. Behavior Research Methods, Instruments and Computers, 17, 217-221.

Carroll, J. B. (1974, May). Psychometric tests as cognitive tasks: A new "structure of intellect" (Tech. Rep. No. 4: ETS-RB-74-16). Washington, DC: Office of Naval Research, Personnel and Training Research Program Office.

Carter, R. C., Kennedy, R. S., & Bittner, A. C., Jr. (1981). Grammatical reasoning: A stable performance yardstick. Human Factors, 23, 587-591.

Carter, R. C., Krause, M., & Harbeson, M. M. (1986). Beware the reliability of slope scores for individuals. Human Factors, 28, 673-683.

Carter, R. C., & Sbisa, H. E. (1982). Human performance tests for repeated measurements; alternate forms of eight tests by computer (Research Rep. No. NBDL-82R003). New Orleans, LA: Naval Biodynamics Laboratory. (NTIS No. AD A115021)

Carter, R. C. & Wolstad, J. C. (1985). Repeated measurements of spatial ability with the Manikin test. Human Factors, 27(2), 209 219.

Detterman, D. K. (1984, August). Computer assisted assessment of cognitive abilities. Paper presented at the 92nd Annual Meeting of the American Psychological Association, Toronto, Canada.

Donders, F. C. Die Schnelligkeit psychischer Processe. Archiv fur Anatomie und Physiologie und Wissenschaftliche Medizin, 1868, 657-681. (Also, Donders, F. C. On the speed of mental processed. (Translated by W. G. Koster) Acta Psychologica, 1969, 30, 412-431.)

Dunlap, W. P., Kennedy, R. S., Harbeson, M. M., & Fowlkes, J. E. (1988, in press). Difficulties with individual difference measures based upon some componential cognitive paradigms. Applied Psychological Measurement.

Ekstrom, R. B., French, J. W., Harman, H. H., & Dermen, D. (1976, August). Manual for kit of factor-referenced cognitive tests (Office of Naval Research Contract No. N000014-71-C-0117). Princeton, NJ: Educational Testing Service.

Essex Corporation (1985). Automated portable test system. Orlando, FL: Brochure.

Englund, C. E., Reeves, D. L., Shingledecker, C. A., Thorne, D. R., Wilson, K. P., & Hegge, F. W. (1987). Unified Tri-Service Cognitive Performance Assessment Battery (UTC-PAB): I. Design and specification of the battery (Rep. No. 87-10). San Diego, CA: Naval Health Research Center.

Fleishman, E. A., & Hempel, W. E., Jr. (1954). Changes in factor structure of a complex psychomotor test as a function of practice. Psychometrika, 19, 239-252.

Fleishman, E. A., & Hempel, W. E., Jr. (1955). The relation between abilities and improvement with practice in a visual discrimination reaction task. Journal of Experimental Psychology, 49, 301-316.

Fleishman, E. A., & Rich, S. (1963). Role of kinesthetic and spatial-visual abilities in perceptual-motor learning. Journal of Experimental Psychology, 66, 6-11.

Guilford, J. P. (1954). Psychometric methods (2nd ed.). New York: McGraw-Hill.

Gullion, C. M., & Eckerman, D. A. (1986). Field testing for neurobehavioral toxicity: Methods and methodological issues. In Z. Annau (Ed.), Neurobehavioral toxicology. Baltimore, MD: John Hopkins.

Hunt, E. B., & Pellegrino, J. (1986). Testing and measures of performance. Proceedings of the 27th Annual Meeting of the Psychonomic Society (pp. 385). New Orleans, LA.

Hunter, D. R. (1975). Development of an enlisted psychomotor/perceptual test battery (AFHRL-TR-75-60). Wright Patterson Air Force Base, OH: Air Force Human Resources Laboratory.

Jerison, H. J. (1955, December). Effect of a combination of noise and fatigue on a complex counting task (WADC TR-55-360). Wright-Patterson Air Force Base, OH: Wrigh Air Development Center, Air Research and Development Command, United States Air Force.

Jones, M. B. (1970a). A two-process theory of individual differences in motor learning. Psychological Review, 77(4), 353-360.

Jones, M. B. (1970b). Rate and terminal processes in skill acquisition. American Journal of Psychology, 83(2), 222-236.

Jones, M. B. (1980). Stabilization and test definition in a performance test battery (Final Rep. Contract N00203-79-M-5089). New Orleans, LA: U.S. Naval Aerospace Medical Research Laboratory. (NTIS No. AD A099987)

Jones, M. B., Kennedy, R. S., & Bittner, A. C., Jr. (1981). A video game for performance testing. American Journal of Psychology, 94, 143-152.

Kennedy, R. S., Carter, R. C., & Bittner, A. C., Jr. (1980). A catalogue of Performance Evaluation Tests for Environmental Research. Proceedings of the 24th Annual Meeting of the Human Factors Society (pp. 344-348). Los Angeles, CA.

Kennedy, R. S., Dunlap, W. P., Jones, M. B., Lane, N. E., & Wilkes, R. L. (1985). Portable human assessment battery: Stability, reliability, factor structure, and correlation with tests of intelligence (Final Rep. NSF/ BNS 85001; also EOTR 85-1). Washington, DC: National Science Foundation. (NTIS No. P888-116645/A03)

Kennedy, R. S., Lane, N. E., & Kuntz, L. A. (1987, August). Surrogate measures: A proposed alternative in human factors assessment of operational measures of performance. Paper presented at the 1st Annual Workshop on Space Operations, Automation & Robotics, Houston, TX: NASA/Johnson Space Center.

Kennedy, R. S., Wilkes, R. L., Dunlap, W. P., & Kuntz, L. A. (1987). Development of an automated performance test system for environmental and behavioral toxicology studies. Perceptual and Motor Skills, 65, 947-962.

Kennedy, R. S., Wilkes, R. L., Kuntz, L. A., & Baltzley, D. R. (1988, October). A menu of self-administered microcomputer based neurotoxicology tests (EOTR 88-10). Orlando, FL: Essex Corporation.

Kennedy, R. S., Wilkes, R. L., Lane, N. E., & Homick, J. L. (1985). Preliminary evaluation of a microbased repeated-measures testing system (Tech. Rep. No. EOTR-85-1). Orlando, FL: Essex Corporation.

Klein, R., & Armitage, R. (1979). Rhythms in human performance: 1 1/2-hour oscillations in cognitive style. Science, 204, 1326-1328.

Krause, M., & Bittner, A. C., Jr. (1982). Repeated measures on a choice reaction time task (Res. Rep. No. NBDL-82R006). New Orleans: Naval Biodynamics Laboratory. (NTIS No. AD A121904)

Krause, M., & Kennedy, R. S. (1980). Performance Evaluation Tests for Environmental Research (PETER): Interference susceptibility test. Proceedings of the 7th Psychology in the DoD Symposium (pp. 459-464). Colorado Springs, CO: USAF Academy.

Naitoh, P. (1982). Chronobiologic approach for optimizing human performance. In F. M. Brown & R. C. Gaeber (Eds.), Rhythmic aspects of behavior (pp. 41-103). Hillsdale, NJ: Eribaum Associates.

NEC Home Electronics (USA), Inc. (1983). NEC PC-8201A users guide. Tokyo: Nippon Electric Co., Ltd.

Phillips, W. A. (1974). On the distinction between sensory storage and short-term visual memory. Perception and Psychophysics, 6, 283-290.

Reeves, D. L., & Thorne, D. R. (1988). Development and application of the Unified Tri-Service Cognitive Assessment Battery within naval aviation. Paper presented at the 59th Annual Scientific Meeting of the Aerospace Medical Association. New Orleans, LA.

Shingledecker, C. A. (1984). A task battery for applied human performance assessment research (Tech. Rep. No. AFAMRL-TR-84). Dayton, OH: Air Force Aerospace Medical Research Laboratory.

Spearman, C. (1904). The proof and measurement of association between two things. American Journal of Psychology, 15, 72-101.

Steinberg, E. P. (1986). Practice for the armed services test. New York, NY: Acco Publishing Co.

Sternberg R. J. (1977). Intelligence, information processing, and analogical reasoning: The componential analysis of human abilitites. Hillsdale, NJ: Lawrence Erlbaum Associates.

The Psychological Corporation (1981). Wechsler adult intelligence scale-revised. New York, NY: Harcourt Brace & Jovanovich.

Thorne, D. R., Genser, S. G., Sing, H. C., & Hegge, F. W. (1985). The Walter Reed Performance Assessment Battery. Neurobehavioral Toxicology & Teratology, 7, 415-418.

Turnage, J. J., Kennedy, R. S., Osteen, M. K. (1987). Repeated-measures analyses of selected psychomotor tests from PAB and APTS: Stability, reliability, and cross-task correlations. Orlando, FL: Essex Corporation.

Underwood, B. J., Boruch, R. F., & Malmi, R. A. (1977, May). The composition of episodic memeory (ONR Contract No. N00014-76-C-0270). Evanston, IL: Northwestern University, (NTIS No. Ad A040696).

Wechsler, D. (1958). Measurement and appraisal of adult intelligence (4th ed.). Baltimore: Williams and Wilkins Company.

Wickens, C. D., Sandry, D., & Vidulich, M. (1983). Compatibility and resource competition between modalities of input, central processing, and output: Testing a model of complex task performance. Human Factors, 25, 227-248.

Wilkes, R. L., Kennedy, R. S., Dunlap, W. P., & Lane, N. E. (1986). Stability, reliability, and cross-mode correlation of tests in a recommended 8-minute performance assessment battery (Tech. Rep. No. EOTR-86-4). Orlando, FL: Essex Corporation.

Wonderlic, C. F. (1983). Wonderlic personnel test. Northfield, IL: E. F. Wonderlic.

Zeider, J. (1987, April). The validity of selection and classification procedures for predicting job performance (IDA Paper P-1977). Alexandria, VA: Institute for Defense Analyses.